

METHOD AND APPARATUS FOR DYNAMICALLY SCHEDULING REQUESTS

Abstract of the Disclosure

Techniques are provided for dynamically scheduling requests in data processing systems in accordance with differentiated service levels. In a first aspect of the invention, a technique for processing a request to at least one server comprises the following steps. A request is received. Then, submission of the request to the at least one server is scheduled based on: (i) a quality-of-service (QoS) class assigned to a client from which the request originated; (ii) a response target associated with the QoS class; and (iii) an estimated response time associated with the at least one server. The technique may further comprise the step of withholding the request from submission to the at least one server when the request originated from a client assigned to a first QoS class to allow a request that originated from a client assigned to a second QoS class to meet a response target associated therewith.